Автономная некоммерческая организация
«Институт Искусственного интеллекта»

121170, Москва, Кутузовский пр-т, 32, к. 1, 4.В.08
E-mail: info@airi.net, Web: airi.net
ОГРН 1207700493978 ИНН 7730261209 КПП 773001001

ОТЗЫВ РЕЦЕНЗЕНТА
на выпускную квалификационную работу

на дипломную работу Савкина Максима Константиновича
на тему: Генерация синтетического датасета для задачи распознавания именованных сущностей с использованием больших языковых моделей

Выпускная работа посвящена генерации синтетических датасетов для задачи распознавания именованных сущностей (NER). Актуальность темы подтверждается тем, что каждая конкретная задача NER требует специализированного обучающего корпуса, однако ручная разметка таких данных затратна по времени и ресурсам, а автоматические методы часто дают низкое качество. Особенно остро эта проблема стоит в области детекции персональной информации, где доступ к реальным данным ограничен из-за юридических и этических причин. Предложенный подход позволяет автоматически генерировать тексты с заданной структурой, что существенно упрощает создание обучающих выборок для различных подзадач NER.

Достоверность результатов подтверждается следующими фактами: предложенная методология генерации синтетических данных формализована и воспроизводима; исходный код и данные доступны в открытом доступе; проведено сравнение с существующими решениями; выполнен анализ сгенерированных данных.

Практическая значимость работы подтверждается популярностью модели показанной в DeepPavlov Demo, активно скачиваемой на регулярной основе. Также значимость подтверждается наличием двух публикаций на конференциях ранга А и А*.

К плюсам работы следует отнести актуальность разработанных датасетов, глубину проведенного анализа и публикацию результатов на международных конференциях. Материал работы ясно изложен. Вместе с тем, работа имеет некоторые недостатки: в экспериментах не приведены стандартные отклонения, не указано сколько раз проводился эксперимент, приводится определение модели трансформер, но нет ссылки на саму статью, diverse beam search, logit suppression также приводятся без ссылок, не обсуждаются ограничения предложенного подхода на доменах, которые могут быть не знакомы для LLM.

Вопросы:

- В моделях для задачи NER часто используется CRF слой, почему он не использовался в данной работе?
- В функции потерь (формула 2.4) нет нормировки на длину текста. Получается, что вклад примеров разной длины во время обучения разный, это сделано намеренно?

- Основной метрикой является F1 на уровне сущностей (span level), но выбранная функция потерь оптимизирует работу модели на уровне токенов (token level). Можно ли подобрать функцию потерь, которая бы учитывала сигнал на уровне сущностей?

В целом, работа М.К. Савкина вносит значимый вклад в область обработки естественного языка и решение задачи распознавания именованных сущностей и обладает практической ценностью.

С учетом высказанных замечаний, можно заключить, что работа М.К. Савкина «Генерация синтетического датасета для задачи распознавания именованных сущностей с использованием больших языковых моделей» является законченной выпускной квалификационной работой и может быть рекомендована к защите с оценкой «отлично».

Рецензент, Куратов Юрий Михайлович
к.ф.-м.н., старший научный сотрудник,
руководитель группы “Модели с памятью”,
лаборатория когнитивных систем искусственного интеллекта,
АНО «Институт искусственного интеллекта»

Подпись рецензента, сотрудника АНО «Институт искусственного интеллекта»,

удостоверяю
HR Директор – начальник отдела
АНО «Институт искусственного интеллекта» Звонарева М.И.